

City University of New York (CUNY)

CUNY Academic Works

Open Educational Resources

Queensborough Community College

2020

Clear-Sighted Statistics: Module 3: Where Do Data Come From? (slides)

Edward Volchok

CUNY Queensborough Community College

[How does access to this work benefit you? Let us know!](#)

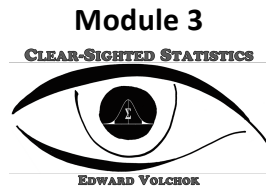
More information about this work at: https://academicworks.cuny.edu/qb_oers/146

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).

Contact: AcademicWorks@cuny.edu

Where Do Data Come From?



Lecture Outcomes

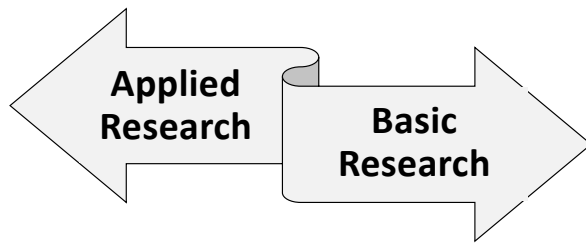
The research process	Primary and Secondary Research	Pure Research versus Applied Research	Types of research: Exploratory, descriptive, causal, and meta-analysis
Define the term big data	Basic sampling techniques	Sampling error	Systematic errors
	Reliability and Validity	Steps to take when reviewing research	

The Research Process

Two broad categories of research

Applied Research	Basic or Pure Research
<ul style="list-style-type: none">• Pragmatic solutions to pressing problems• Research to help decision-makers	<ul style="list-style-type: none">• Advance knowledge, not than solve immediate problems

A continuum, not binary

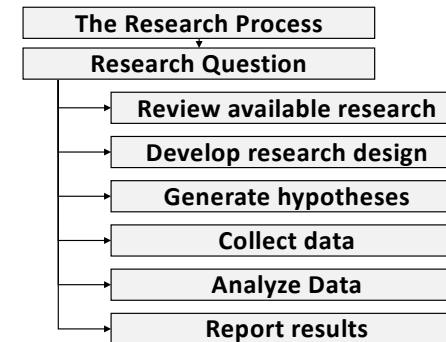


© 2019

Clear-Sighted Statistics

4

The multi-stage research process



© 2019

Clear-Sighted Statistics

5

Typical research questions

How do Facebook users feel about the privacy of their data?

Which candidates for President of the United States get increased Google searches after a presidential debate and who is searching?

What behavioral, biological, pharmacological, and treatment factors contribute to better blood glucose control for Type I diabetics?

What factors have contributed to the decline of Cable TV subscriptions?

What factors are associated with higher graduation rates at community colleges and four-year colleges?

© 2019

Clear-Sighted Statistics

6

Review available research

Secondary Research

- Previously conducted research that's available to public
- Published studies
- Government reports

Primary Research

- Proprietary research, not available to public

© 2019

Clear-Sighted Statistics

7

Secondary research is useful

Provides background information

Can help researchers refine their questions

Might answer the researchers' question

Might alert researchers about problems they need to avoid

Might help researchers decide on the most appropriate research methods to use

Might provide information that will help researchers with their sampling



8

Questions to ask about secondary research...

Who conducted the research?

For what purpose was the research done?

How were the data collected?

When was the data collected?

What is included in the data and what is not?

Are the findings consistent with other research?



9

Limitations of Secondary Research

Appropriate secondary data may not be available

May be out-of-date

May lack relevance for the problem being investigated

May be inaccurate

While useful, they may not adequately address the problem under investigation



10

Is primary research necessary?

Does the secondary research answer the research question?

If yes, primary research may not be needed

If no, primary research may be needed when the research question is important and there is enough time and money to conduct that research



11

Research objectives must be SMART

Specific

Measurable

Achievable

Relevant

Timed

© 0 9 1

12

12

Four basics types of research

Exploratory

Descriptive

Causal (Experiment or Random Controlled Tests)

Meta-Analysis

© 0 9 1

13

13

Exploratory Research

Preliminary study designed to get a deeper understanding of a problem

No quantitative information is obtained

No hypotheses are tested

Focus group, in-depth interviews

© 0 9 1

14

14

Descriptive Research

Addresses: Who? What? Where? When? Why?

Can be qualitative or quantitative

Examples: Observational Studies, Case Studies, and Surveys

© 0 9 1

15

15

Observational Studies

Prospective

- Data collected as events occur

Retrospective

- Data collected after events have occurred

CC BY-NC-SA

16

16

Case Studies

Detailed investigations of one or two examples of an issue in hopes of finding lessons for all similar cases

Quantitative and qualitative data are collected

CC BY-NC-SA

17

17

Surveys

Respondents provide data by answering a questionnaire

The wording of the questionnaire is extremely important (to be discussed shortly)

CC BY-NC-SA

18

18

Causal Research

Addresses the “why” question

Controlled experiments used to determine cause and effect

Other forms of research may establish causality

Cause must precede the effect

Concomitant variations between cause and effect

CC BY-NC-SA

19

19

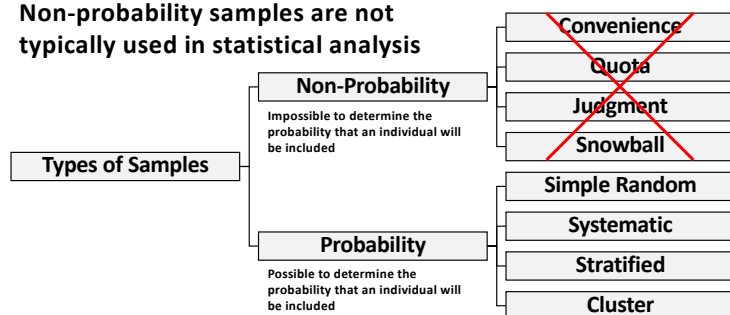
Meta-Analysis

Statistical techniques used to develop quantitative analyses of multiple studies on the same or similar subjects

Sampling Techniques

Sourcing data with samples

Non-probability samples are not typically used in statistical analysis



Simple Random Samples

Sample is chosen on the basis of chance

Each item in the population has the same chance of being included

Examples: Mega Millions and Powerball lotteries

Probability of a population element being selected

$$\frac{\text{Sample Size}}{\text{Population Size}} = \frac{50}{15,000} = 0.0067 = 0.67\%$$

Advantages & Disadvantages of Simple Random Samples

Cost & Use

- High cost
- Most common use is random dialing

Advantages

- Minimal advanced knowledge of the population is needed
- Easy to analyze data and computer error

Disadvantages

- Requires a sampling frame
- Does not use established knowledge of the population
- Larger sample error per sample size
- Respondents may be widely dispersed.

Systematic Random Samples

Population elements are ordered

Random starting point is selected, and then every k^{th} member of the population is selected

When physical order of population is related to characteristics, this method should not be used

Advantages & Disadvantages of Systematic Random Samples

Cost & Use

- Moderate cost
- Moderate use

Advantages

- Simple to draw
- Easy to check

Disadvantages

- Order on the population may introduce errors

Stratified Samples

Population is first divided into subgroups (strata) based on important characteristics

Sample is then selected from each stratum

Ensures proportionate representation of each strata

© 2019

Clear-Sighted Statistics

28

28

Stratified Sample: Example

Study advertising spending for 400 large corporations (Sample Size = 50)

Question: Do firms with high Return on Equity invest more in advertising?

To make sample representative, companies are grouped by Return on Equity

Stratum	Return on Equity	Number	Relative Frequency	Number Sampled
1.00	30% or more	10.00	2.50%	1.00
2.00	20% < 30%	77.00	19.25%	10.00
3.00	10% < 20%	192.00	48.00%	24.00
4.00	0% < 10%	109.00	27.25%	14.00
5.00	Deficit	12.00	3.00%	1.00
Total		400	100.00%	50

© 2019

Clear-Sighted Statistics

29

29

Advantages & Disadvantages of Stratified Samples

Cost & Use

- High cost
- Moderate use

Advantages

- Ensures representation of all groups in the sample
- Allows for estimates of the population and comparisons of the strata
- Reduces variability for the sample size

Disadvantages

- Requires accurate information of the proportion of each stratum
- If stratified lists are not available, it can be costly to develop these lists

© 2019

Clear-Sighted Statistics

30

30

Cluster Samples

Population is first divided into sub-groups (clusters), then samples are randomly selected from the clusters

Example: Determine citizens' views on environmental protection

Using cluster sampling by dividing population into smaller units (clusters)

Divide country into clusters and sample from a selected number randomly selected clusters

© 2019

Clear-Sighted Statistics

31

31

Advantages & Disadvantages of Cluster Samples

Cost & Use

- High cost
- Moderate use

Advantages

- Ensures representation of all groups in the sample
- Allows for estimates and comparisons of the strata
- Reduces variability for the sample size

Disadvantages

- Requires accurate information of the proportion of each stratum
- If stratified lists are not available, it can be costly to develop these lists

Random Sampling Error

Random Sampling Errors

Sample Statistic \neq Population Parameter

Not due to human error

All samples have some degree of sampling error

Sampling error is a major issue in inferential statistics

Systematic Errors

Systematic Errors

Human errors in the design and execution of the research

Sample Design Errors: Problems with how the sampling was conducted

Measurement Errors: Errors made when the data are collected (Can happen with samples and censuses)



Clear-Sighted Statistics

36

36

Sample Design Errors: Frame Error

Sample frame is a list of all the elements in a population

Frame error occurs when the sample frames are inaccurate

Frame error biases research results



Clear-Sighted Statistics

37

37

Sample Design Errors: Population Specification Error

Errors that result from incorrectly defining the population



Clear-Sighted Statistics

38

38

Sample Design Errors: Selection Error

Even with an accurate sample frame and properly defined population, selection error may happen

Selection errors occur when the sampling procedures are not properly followed or the procedures are poorly thought out

Example: Interviewer avoids interviewing certain types of people



Clear-Sighted Statistics

39

39

Sample Design Errors: Measurement Error – Surrogate Information Error

Errors due to an inconsistency between the information sought and the information needed to solve a problem

These errors arise when the researcher fails to understand how respondents view the questions

© 0 0 0 1

Clear-Sighted Statistics

40

40

Sample Design Errors: Measurement Error – Possessing Error

Errors that occur after the data have been collected

These are errors in coding, transcribing, assigning weights to the data as well as use of inappropriate statistical techniques

© 0 0 0 1

Clear-Sighted Statistics

41

41

Measurement Errors: Interviewer Error

Interviewer consciously or unconsciously influences respondents' answers

Respondents' reaction to the interviewer may introduce bias

Interviewer Error introduce bias

© 0 0 0 1

Clear-Sighted Statistics

42

42

Measurement Errors: Instrument or Questionnaire Bias

Leading Questions: Embed the answer in the question and bias results

How dumb was President Obama's policy on North Korea?

© 0 0 0 1

Clear-Sighted Statistics

43

43

Measurement Errors: Instrument or Questionnaire Bias

Loaded Question: Contain unjustified assumptions that skew responses

Do you think the liberal media push fake news to undermine President Trump?

CC BY-NC-SA

44

44

Measurement Errors: Instrument or Questionnaire Bias

Double Barreled: Two questions embedded in a single question

Would you vote for a candidate who supports cutting spending on education and health care?

CC BY-NC-SA

45

45

Measurement Errors: Response Bias

Acquiescence Bias: Tendency of respondents to agree with all questions, or to answer all question with positive responses

Extremity Bias: Tendency of respondents to use extremely positive or negative responses

Auspices Bias: Tendency of respondents to be influenced by the study's sponsor

CC BY-NC-SA

46

46

Measurement Errors: Non-Response Bias

Bias that occurs when respondents fail to participate fully in the survey

Theoretical difference between a perfect survey in which everyone participates and the actual survey

Self-Selection Bias: Respondents tend to have strongly positive or negative views

CC BY-NC-SA

47

47

Experimental Errors:

The reactive effects that happen when respondents know they are participating in a study

Hawthorne Effect: Participants do not behave normally because they know they are in a study

Placebo Effect: All participants show positive results regardless of what group they are assigned

John Henry Effect: Participants in the control group work harder

Reliability and Validity

Reliability and Validity

Reliability refers to the consistency of the questionnaire

Validity refers to the accuracy of the measurements

Reliability and Validity

Poor Validity, Good Reliability



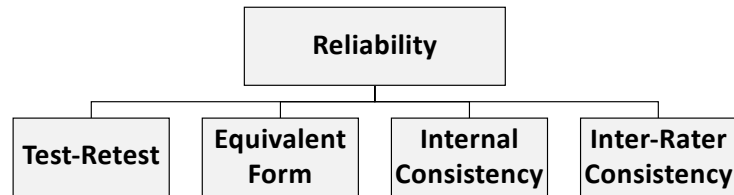
Poor Validity, Poor Reliability



Good Validity, Good Reliability



Types of Reliability



© ⓘ ⓘ ⓘ ⓘ ⓘ

52

52

Test-Retest Reliability

Respondents are given the survey at two different times under similar conditions

When there are few differences, the questionnaire is said to be reliable

© ⓘ ⓘ ⓘ ⓘ ⓘ

53

53

Problems with Test-Retest Reliability

It may be difficult or impossible to get respondents to complete the questionnaire twice

The administration of the first survey may cause a respondent's answers to change

Extraneous factors in the respondents' environment may cause the measurements to change

© ⓘ ⓘ ⓘ ⓘ ⓘ

54

54

Equivalent Form Reliability

Similar surveys are used to determine whether they yield similar results

When results are similar, the surveys are considered reliable

© ⓘ ⓘ ⓘ ⓘ ⓘ

55

55

Internal Consistency Reliability

Questionnaire returns similar measurements when given to different samples during the same time period

Split ballot technique: Questions that probe the same construct are divided into two groups of equal numbers of questions

Chronbach's Alpha: Calculates the mean reliability scores for all possible ways of splitting the questions

© 0 0 0 1

Clear-Sighted Statistics

56

Interpreting Chronbach's Alpha, α

Chronbach's Alpha Internal Consistency

$\alpha \geq 0.9$	Excellent
$0.8 \geq \alpha < 0.9$	Good
$0.7 \geq \alpha < 0.8$	Acceptable
$0.6 \geq \alpha < 0.7$	Questionable
$0.5 \geq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

© 0 0 0 1

Clear-Sighted Statistics

57

Inter-Rater Consistency Reliability

How independent observers compare their assessments when they observe the same behavior

Example: Quantitative scores of Olympic judges of a gymnastics competition are consistent

© 0 0 0 1

Clear-Sighted Statistics

58

Validity

Validity deals with how accurately a questionnaire measures a concept or "construct"

© 0 0 0 1

Clear-Sighted Statistics

59

Researchers use two definitions for constructs

Theoretical Definition

- States the essential meaning of the construct

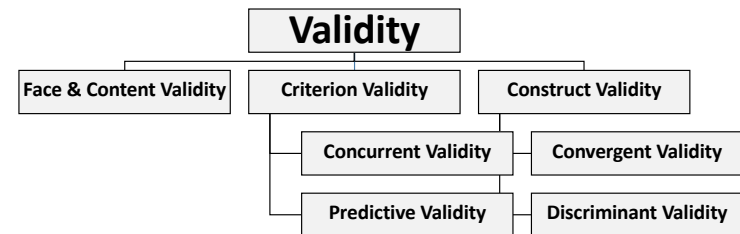
Operational Definition

- A description of the observable features of a construct that will be measured

© 0 9 1

60

Types Validity



© 0 9 1

61

Face and Content Validity

Non-quantitative forms of validity

Face Validity: Degree to which non-experts *judge* the results to be accurate

Content Validity: Degree to which experts *judge* the results to be accurate

© 0 9 1

62

Criterion Validity

Degree to which the questionnaire can predict the designated criterion

Concurrent Validity

Predictive Validity

© 0 9 1

63

Criterion Validity: Concurrent Validity

The degree to which one variable measured can be predicted by the instrument concurrently—at the same time—with another variable of interest

If we have a predictive relationship, we say that these variables concur

© 0 9 1

Clear-Sighted Statistics

64

64

Criterion Validity: Predictive Validity

The degree to which the future level of the criterion variable can be predicted by the instrument

SAT scores would have predictive validity if they predict students' academic performance in college

© 0 9 1

Clear-Sighted Statistics

65

65

Construct Validity

The degree to which the questionnaire accurately measures the construct

Convergent Validity

Discriminant Validity

© 0 9 1

Clear-Sighted Statistics

66

66

Construct Validity: Convergent Validity

The strength of the association among different questions or questionnaires that purport to measure the same construct

When the measures for these constructs are similar—when they converge— we have convergent validity

© 0 9 1

Clear-Sighted Statistics

67

67

Construct Validity: Discriminant Validity

The lack of association among constructs that are supposed to be different

When test results show that measures that are supposed to be dissimilar are, in fact, dissimilar

10 questions to ask when reviewing research

Questions skeptics ask when reviewing research

Have the study's
goal been
identified?

Has the study's
methodology
been described?

Who funded
the study?

Have the results
been replicated?

Is the sample size
sufficient?

Are the variables
adequately
defined and
measured?

Is the
questionnaire
reliable and
valid?

Have potential
sources of
systematic errors
been identified?

Are the tables
and charts
misleading?

Are there
alternate
explanations?


The
End




Clear-Sighted Statistics

Except where otherwise noted *Clear-Sighted Statistics* is licensed under a Creative Commons License. You are free to share derivatives of this work for non-commercial purposes only. Please attribute this work to Edward Volchok.

CLEAR-SIGHTED STATISTICS



EDWARD VOLCHOK

 72